Molecular Microbiology (1997) 25(4), 619–637

Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea

Eugene V. Koonin,[1*] Arcady R. Mushegian,[1†] Michael Y. Galperin[1] and D. Roland Walker[1,2]
[1]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.*
[2]*Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA.*

Summary

Protein sequences encoded in three complete bacterial genomes, those of *Haemophilus influenzae*, *Mycoplasma genitalium* and *Synechocystis* sp., and the first available archaeal genome sequence, that of *Methanococcus jannaschii*, were analysed using the BLAST2 algorithm and methods for amino acid motif detection. Between 75% and 90% of the predicted proteins encoded in each of the bacterial genomes and 73% of the *M. jannaschii* proteins showed significant sequence similarity to proteins from other species. The fraction of bacterial and archaeal proteins containing regions conserved over long phylogenetic distances is nearly the same and close to 70%. Functions of 70–85% of the bacterial proteins and about 70% of the archaeal proteins were predicted with varying precision. This contrasts with the previous report that more than half of the archaeal proteins have no homologues and shows that, with more sensitive methods and detailed analysis of conserved motifs, archaeal genomes become as amenable to meaningful interpretation by computer as bacterial genomes. The analysis of conserved motifs resulted in the prediction of a number of previously undetected functions of bacterial and archaeal proteins and in the identification of novel protein families. In spite of the generally high conservation of protein sequences, orthologues of 25% or less of the *M. jannaschii* genes were detected in each individual completely sequenced genome, supporting the uniqueness of archaea as a distinct domain of life. About 53% of the *M. jannaschii* proteins belong to families of paralogues, a fraction similar to that in bacteria with larger genomes, such as *Synechocystis* sp. and *Escherichia coli*, but higher than that in *H. influenzae*, which has approximately the same number of genes as *M. jannaschii*. Certain groups of proteins, e.g. molecular chaperones and DNA repair enzymes, thought to be ubiquitous and represented in the minimal gene set derived by bacterial genome comparison, are missing in *M. jannaschii*, indicating massive non-orthologous displacement of genes responsible for essential functions. An unexpectedly large fraction of the *M. jannaschii* gene products, 44%, shows significantly higher similarity to bacterial than to eukaryotic proteins, compared with 13% that have eukaryotic proteins as their closest homologues (the rest of the proteins show approximately the same level of similarity to bacterial and eukaryotic homologues or have no homologues). Proteins involved in translation, transcription, replication and protein secretion are most closely related to eukaryotic proteins, whereas metabolic enzymes, metabolite uptake systems, enzymes for cell wall biosynthesis and many uncharacterized proteins appear to be 'bacterial'. A similar prevalence of proteins of apparent bacterial origin was observed among the currently available sequences from the distantly related archaeal genus, *Sulfolobus*. It is likely that the evolution of archaea included at least one major merger between ancestral cells from the bacterial lineage and the lineage leading to the eukaryotic nucleocytoplasm.

Introduction

Microbiology has entered a new era that is marked by the availability of complete genome sequences of bacteria, archaea and unicellular eukaryotes for comparative analysis. At the time of writing (November, 1996), the completely sequenced genomes include those of four bacteria, namely *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Mycoplasma genitalium* (Fraser *et al.*, 1995),

Received 6 February, 1997; revised 6 June, 1997; accepted 11 June, 1997. †Present address: Sequana Therapeutics, Inc., 11099 North Torrey Pines Rd., La Jolla, CA 92037, USA. *For correspondence. E-mail koonin@ncbi.nlm.nih.gov; Tel. (301) 496 2477; Fax (301) 480 9241.

Mycoplasma pneumoniae (Himmelreich *et al.*, 1996) and *Synechosystis* sp. (Kaneko *et al.*, 1996), one archaeon, *Methanococcus jannaschii* (Bult *et al.*, 1996) and one unicellular eukaryote, the yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996). Thus, representative genome sequences from each of the three primary domains of life (Woese *et al.*, 1990) are already available, creating an opportunity for the reconstruction of the genome organization of the last common ancestor of all modern life forms. In the course of such a reconstruction, traces of major evolutionary events that shaped the genomes of bacteria, archaea and eukaryotes may be discovered.

Analysis of complete genomes allows us to address biologically important problems that were previously not tractable (Fleischmann *et al.*, 1995; Koonin *et al.*, 1996a; Koonin and Mushegian, 1996). In particular, only the comparison of complete gene sets enables the matching of all known biochemical pathways with specific genes. It is becoming possible to ascertain, given sufficient sensitivity of the methods used for sequence comparison, that certain protein families are *not* encoded in a given genome and to seek alternative candidates for essential roles among the gene products. Comparison of complete gene sets should also allow researchers to define molecular correlates of the unique lifestyles of different species.

Complete genome sequences will yield biological insights only if the functions of the gene products are predicted in as much detail as possible. Compared with original, conservative database searches, the use of additional computer approaches, specifically careful analysis of relatively weak sequence similarities by multiple alignment construction and motif detection, tends to produce a wealth of information on protein functions and relationships (Bork *et al.*, 1992; 1995; Koonin *et al.*, 1994; Ouzounis *et al.*, 1995; Koonin *et al.*, 1995, 1996b; Tatusov *et al.*, 1996). Based on the detailed prediction of protein functions, it becomes possible to reconstruct, at least in its general features, the biochemistry of a poorly studied bacterial species (Fleischmann *et al.*, 1995; Tatusov *et al.*, 1996). Comparison of the complete sets of proteins encoded by the genomes of phylogenetically distant species can be used to predict functions and genes that are essential for cellular life (Fraser *et al.*, 1995; Mushegian and Koonin, 1996a; Koonin and Mushegian, 1996).

The first sequenced archaeal genome, that of *M. jannaschii*, appears to be particularly interesting and unusual compared with the bacterial genomes. The results of the original sequence analysis indicated that 56% of the proteins encoded in this genome showed no sequence similarity to any available protein sequences from other species (Bult *et al.*, 1996). Clearly, this is caused in part by the paucity of sequence information from other archaea contained in current databases. With several complete genomes now available and genomes of other, diverse species covered extensively, the great majority of bacterial and eukaryotic protein families are probably represented in the current databases. Thus, it is important to find out whether or not a larger fraction of the *M. jannaschii* protein sequences can be matched with sequences from the other two domains through the use of more sensitive computer methods.

Here, we present the results of comparative analysis of the protein sequences from *M. jannaschii* and three complete bacterial genomes; comparisons with the yeast protein sequences were also made where this was considered important, although detailed analysis of the yeast genome is outside the scope of this work. Our goal was to detect sequence similarities, including relatively weak but apparently biologically relevant ones, and to use them for functional prediction to the maximum extent possible, with the aim of revealing common and distinctive features of archaeal and bacterial genomes.

Results and discussion

About 70% of the proteins encoded in each bacterial or archaeal genome contain highly conserved regions

It has been observed previously that the great majority of proteins encoded in a bacterial genome, e.g. those of *Mycoplasma capricolum* (Bork *et al.*, 1995), *E. coli* (Koonin *et al.*, 1995, 1996b) and *H. influenzae* (Fleischmann *et al.*, 1995; Casari *et al.*, 1995; Tatusov *et al.*, 1996), show sequence similarity to proteins contained in databases and, more importantly, that for most of the bacterial proteins, the observed conservation is not limited to closely related species (Koonin *et al.*, 1995; Tatusov *et al.*, 1996). Here, we have evaluated the protein sequence conservation for three complete bacterial genomes and one archaeal genome in a single computer experiment. The complete sets of protein sequences encoded in the genomes of *H. influenzae*, *M. genitalium*, *Synechocystis* sp. and *M. jannaschii* were compared with the protein sequence databases using the WUBLASTP program based on the BLAST2 algorithm, and the search output was further analysed for conserved motifs as described under *Experimental procedures*. This analysis revealed very similar distributions of protein sequence conservation for the bacterial and archaeal genomes. Sequence similarity that we interpreted as biologically relevant, based on statistically significant alignments and motif conservation, was detected for 73% of the *M. jannaschii* gene products and for 75–90% of the gene products in each of the three bacterial species (Fig. 1). Only 5% of the *M. jannaschii* proteins are conserved exclusively within archaea, whereas the remaining 68% have either bacterial or eukaryotic homologues, or both (Fig. 1). Thus, the fraction of proteins that contain regions conserved over large evolutionary distances is nearly constant at about 70% in the three

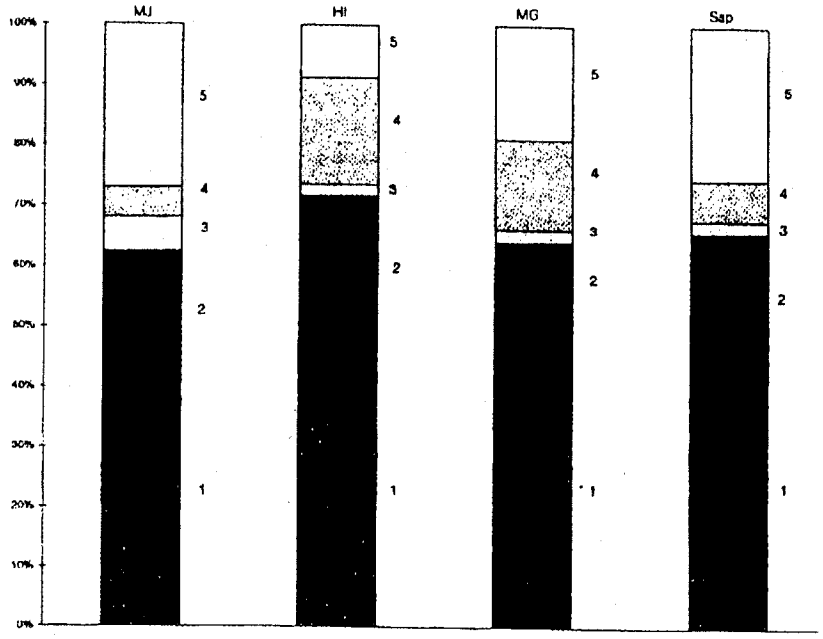

Fig. 1. Sequence conservation among bacterial and archaeal protein sequences. *M. jannaschii*: 1, similarity to both bacterial and eukaryotic proteins; 2, similarity to bacterial proteins only; 3, similarity to eukaryotic proteins only; 4, similarity only to proteins from other archaea; 5, no similarity to proteins from other species. Bacteria: 1, similarity to both eukaryotic or archaeal proteins and proteins from distantly related bacteria; 2, similarity to proteins from distantly related bacteria; 3, similarity to eukaryotic proteins only; 4, similarity to proteins from closely related bacteria only; 5, no similarity to proteins form other species. 'Closely related bacteria' were defined as Proteobacteria for *H. influenzae*, low G+C Gram-positive bacteria for *M. genitalium* and Cyanobacteria for *Synechocystis* sp.

completely sequenced bacterial genomes and the first available archaeal genome (Fig. 1). It is expected that, with the growth of the archaeal sequence data set, the percentage of the *M. jannaschii* protein sequences that have detectable homologues in other species will reach the values observed for bacteria.

On average, sequence similarity to the most closely related sequence in the database was considerably lower for *M. jannaschii* proteins compared with bacterial proteins; for *M. jannaschii*, the median score produced by BLASTP corresponded to a marginal statistical significance, whereas, for each of the bacteria, this score was highly significant (Table 1); hence, the greater role accorded to the more sensitive and selective BLAST2 algorithm as well as methods for motif analysis in the detection of homologous relationships for the archaeal proteins. Compared with the original studies, the increase in the detection of homologues for bacterial proteins was relatively modest but, in the case of *M. jannaschii*, the step up from 44% of gene products with detectable sequence conservation (Bult *et al.*, 1996) to 73% significantly affects our view of the genome.

The majority of bacterial and archaeal proteins are amenable to functional characterization by sequence analysis

Sequence similarity analysis allowed functional prediction for the majority of bacterial and archaeal gene products (Fig. 2). The fraction of proteins for which only a general functional prediction could be made, e.g. of a particular enzymatic or binding activity, but for which attribution of a specific physiological role was not feasible, was significantly greater for *M. jannaschii* than for *H. influenzae* and *M. genitalium*, but very similar to that for *Synechocystis* sp. (Fig. 2). Presumably, this reflects the limitations of the current knowledge of archaeal and cyanobacterial biochemistry. On many occasions, a general indication of protein function is possible even in the absence of detectable sequence similarity, particularly via the identification of signal peptides and transmembrane domains. Only a small fraction of predicted gene products in each of the complete genomes remains totally uncharacterized after detailed sequence analysis (Fig. 2).

A comparison of the distribution of proteins with predicted function by functional classes reveals both common

Table 1. Best database search scores for archaeal and bacterial protein sequences.

Species	Average/median highest score	
	BLASTP	WUBLASTP (BLAST2)
M. jannaschii	174/83[a]	309/148[b]
H. influenzae	586/415	790/611
M. genitalium	312/208	505/317
Synechocystis sp.	279/134	411/216

a. For an average-sized protein (about 300 amino acid residues), a score of 83 approximately corresponds to a *P*-value of 0.01 and is not necessarily indicative of biological relevance of the alignment.
b. For an average-sized protein (about 300 amino acid residues), a score of 148 approximately corresponds to a *P*-value $< 10^{-7}$, which typically indicates a biologically relevant alignment (except for proteins with a highly biased composition).

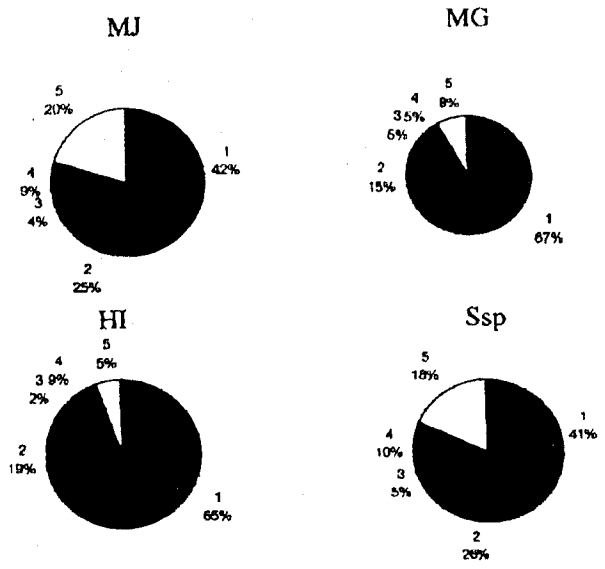

Fig. 2. Functional prediction for bacterial and archaeal proteins. 1, specific prediction including assignment to a functional class; 2, general functional prediction (e.g. enzymatic or binding activity) based on sequence similarity; 3, general functional prediction based on the analysis of structural features (e.g. signal peptides or transmembrane domains); 4, no functional prediction in spite of detected sequence conservation; 5, no prediction.

and distinctive features between *M. jannaschii* and bacteria (Table 2). The fraction of proteins in several categories, namely translation, transcription, replication, energy conversion, cofactor metabolism and inorganic ion transport, is similar for bacteria and *M. jannaschii*, whereas other categories seem to be underrepresented in *M. jannaschii* (Table 2). The latter observation, however, may largely be accounted for by the current insufficient knowledge of archaeal biochemistry. This situation is likely to change with further progress in the study of archaeal genomes and metabolism, perhaps resulting in a distribution of proteins by functions more closely resembling that in bacteria.

Novel findings in the M. genitalium *genome: automatic prediction of protein functions in complete genomes remains elusive*

An automatic computer system called GeneQuiz (Scharf *et al.*, 1994) has recently been used for rapid reanalysis of the *M. genitalium* genome sequence (Ouzounis *et al.*, 1996). It has been indicated that the GeneQuiz results were corroborated by detailed manual analysis and could be considered an important benchmark for assessing the completeness and accuracy of genome analysis (Ouzounis *et al.*, 1996). If shown to be robust, this system would indeed represent major progress in comparative genomics. Since an analysis of the *M. genitalium* protein sequences was an integral part of our project on bacterial and archaeal genome analysis, we compared our results with those of the GeneQuiz.

Our analysis of the *M. genitalium* gene products resulted in 53 sequence similarity-based functional predictions that have not been reported originally (http://://www.ncbi.nlm.nih.gov/Complete_Genomes/Mgen/Novel_Findings). All of these functional assignments were supported by statistically significant sequence similarity and/or motif conservation. In most cases, sequence similarity to functionally uncharacterized proteins in the databases has

Table 2. Functional classification of proteins encoded in complete archaeal and bacterial genomes.

Functional category[a]	Number of proteins			
	M. jannaschii	*H. influenzae*	*M. genitalium*	*Synechocystis* sp.
Amino acid metabolism and transport	102 (5.9%)	162 (9.5%)	17 (3.8%)	158 (5.0%)
Replication, recombination and repair	87 (5.0%)	110 (6.4%)	37 (7.9%)	82 (2.6%)
Transcription	22 (1.3%)	30 (1.8%)	12 (2.8%)	26 (0.8%)
Energy conversion	162 (9.4%)	141 (8.3%)	37 (7.9%)	193 (6.1%)
mRNA translation and ribosome biogenesis	114 (6.6%)	125 (7.3%)	97 (20.7%)	118 (3.7%)
Outer membrane and cell wall	36 (2.1%)	105 (6.2%)	9 (1.9%)	131 (4.1%)
Carbohydrate metabolism and transport	10 (0.6%)	80 (4.8%)	14 (3%)	110 (3.5%)
Nucleotide metabolism and transport	33 (1.9%)	73 (4.3%)	28 (6%)	61 (1.9%)
Cofactor metabolism	89 (5.1%)	70 (4.0%)	8 (1.7%)	107 (3.4%)
Chaperones	12 (0.7%)	53 (3.1%)	15 (3.3%)	68 (2.1%)
Inorganic ion transport	48 (2.8%)	52 (3.0%)	10 (2.1%)	115 (3.6%)
Lipid metabolism	13 (0.7%)	40 (2.3%)	9 (1.9%)	62 (2.0%)
Secretion	23 (1.3%)	35 (2.1%)	20 (4.3%)	45 (1.4%)
General prediction only	494 (28.5%)	360 (19.4%)	95 (20.1%)	974 (30.7%)
Total predicted function	1235 (71.3%)	1457 (85.6%)	408 (87.2%)	2254 (71.1%)

a. As previously (Tatusov *et al.*, 1996; Koonin and Mushegian, 1996), genes coding for proteins implicated in the membrane transport of a particular class of metabolites (e.g. nucleotide components or amino acids) and in the expression regulation of a given functional class of genes were included in the respective class.

been detected previously, but functional prediction became possible only upon more detailed motif analysis.

From the six GeneQuiz predictions considered most interesting and specifically discussed by Ouzounis *et al.* (1996), only two (MG333 and MG385, both phosphodiesterases) could be confirmed. The remaining four cases illustrate problems with the GeneQuiz analysis. The MG123 gene product, identified as an arginine deiminase by Ouzounis *et al.* (1996), cannot possess this enzymatic activity for several reasons. The observed similarity to arginine deiminase from *Mycoplasma arginini* is not highly significant statistically (*P*-value of 0.04) and is much lower than that between arginine deiminases from other *Mycoplasma* species and distantly related (Gram-negative) bacteria. Furthermore, the region of similarity does not include the sequence motifs conserved in the arginine deiminases and does not occupy a similar location in MG123 and arginine deiminases. Most importantly, MG123 is predicted to consist mostly of non-globular domains and is therefore unlikely to possess any enzymatic activity. Thus, MG123 is not the second enzyme of amino acid metabolism in *M. genitalium*. The first enzyme in this category, serine hydroxymethyltransferase (MG394), is likely to be involved only in folate metabolism (Mushegian and Koonin, 1996a); amino acid metabolism may not be represented in *M. genitalium* at all.

The similarity between MG237 and isoleucyl-tRNA synthetases could not be confirmed in our analysis, and the origin of this observation remains uncertain to us. The case of MG449 and phenylalanyl-tRNA synthetase illustrates a case when the reported similarity is valid but the implications are not straightforward. The protein in question is indeed highly similar to the N-terminal region of Phe-tRNA synthetase from *M. genitalium* and other bacteria. Inspection of the three-dimensional structure of the *Thermus thermophilus* Phe-RS (Mosyak *et al.*, 1995) indicates, however, that this region folds into a distinct domain, which shows sequence similarity to a variety of small proteins and domains of larger proteins including, among others, bacterial Met-tRNA synthetases, yeast quadruplex DNA-binding protein and human cytokine EMAP (Frantz and Gilbert, 1995; Koonin *et al.*, 1996b; Simos *et al.*, 1996; E. V. Koonin and A. G. Murzin, unpublished observations). The common function of all these domains remains unclear but obviously has little relation to the enzymatic activity of Phe-RS.

In the last of the six examples, Ouzounis *et al.* (1996) characterize the MG468 gene product as 'DNA polymerase I' noting that the sequence is identical to that of MG262, and the relationship between the two is unclear. In fact, the available genome sequence of *M. genitalium* encodes only one of these proteins, namely MG262. The other one seems to be a fictitious 'duplication' introduced in the course of the original annotation. Furthermore, MG262 is not a DNA polymerase, it is a homologue of the N-terminal, 5′-3′ exonuclease domain of DNA polymerase and can be confidently predicted to possess exonuclease activity (Koonin and Bork, 1996; Himmelreich *et al.*, 1996).

Altogether, of the 21 predictions of novel protein functions made by GeneQuiz, only eight could be fully corroborated (E. V. Koonin, unpublished observations). In contrast, most of the predictions produced by our present analysis have not been reported by GeneQuiz. It appears, therefore, that currently available automatic systems for genome analysis are still far from perfect, and expert evaluation of functional predictions remains the limiting step.

Novel findings in the Methanococcus jannaschii *genome: filling in major gaps in cell physiology and predicting enzymatic activities with as yet unknown cellular role*

The analysis of the protein sequences encoded in the *M. jannaschii* genome resulted in functional prediction for 382 gene products based on previously undetected sequence conservation; a selection of such findings is shown in Table 3. Some of the new functional predictions fill important gaps in the original identifications. For example, it has been indicated that *M. jannaschii* encodes aminoacyl-tRNA synthetases (aaRSases) for only 16 amino acids, with no enzymes identified for glutamine, asparagine, cysteine and lysine (Bult *et al.*, 1996). Asparaginyl-tRNA and glutaminyl-tRNA are thought to be formed via transamidation of the aspartyl-tRNA and glutamyl-tRNA, respectively, a mechanism previously identified in Gram-positive bacteria (Strauch *et al.*, 1988) and in archaea (Curnow *et al.*, 1996). The mechanism of cysteine and lysine activation for incorporation into protein has remained unknown. Our analysis revealed a protein, MJ0539, that, while only distantly related to cysteine aaRSases, contains the principal conserved motifs typical of the aaRSase class I (Eriani *et al.*, 1990). We predict that this protein is responsible for cysteine activation in *M. jannaschii*.

Furthermore, a duplication of the alpha-chain of the phenylalanine aaRSase was detected (Table 3). One of the two diverged copies, namely MJ0487, shows significant similarity to lysine aaRSases from several bacterial species and may be predicted to catalyse Lys-tRNA formation, thus completing the repertoire of aaRSases for *M. jannaschii*.

Most of the novel functional predictions stem from the identification of relatively weak sequence similarities combined with the detection of known or novel conserved motifs. Certain classes of proteins, e.g. ATPases, proteins containing the helix–turn–helix DNA-binding domain and SAM-dependent methyltransferases (see Table 5), which are best recognized by motif analysis, were significantly underpredicted in the original report on

Table 3. Novel functional predictions for *M. jannaschii* gene products (examples).

MJ no.	Paralogues	Best functionally relevant hit from other species, *P*-value; % identity/alignment length; conserved motifs	Predicted function/activity and comments
MJ0050	None	gil1230658 (*S. cerevisiae*); 2.0e-19; 22%/395	Glutamate or histidine decarboxylase
MJ0052	None	gil1221579(*H. influenzae*); 0.047; 27%/95; a catalytic motif conserved in sulphurtransferases and phosphatases (E. V. Koonin, unpublished observations)	Sulphurtransferase (rhodanese homologue) probably involved in cysteine biosynthesis (Table 9)
MJ0109	MJ0917	MYOP_BOVIN; 8.0e-19; 29%/244	Inositol monophosphatase
MJ0137	MJ0651, MJ1495	SOHB_HAEIN; 0.0013; 25%/178; conserved motif around the predicted catalytic serine (E. V. Koonin, unpublished observations)	Periplasmic serine protease
MJ0165	MJ0616	PUR6_METTH; 2.1e-06; 31%/128	Phosphosibosylaminoimidazole (AIR) carboxylase; compared with homologues from other species and with MJ0616, contains an additional, uncharacterized N-terminal domain
MJ0215	MJECL20 (nearly identical gene on an extrachromosomal element)	gil1304389 (rabbit); 0.016; 29%/104; motifs conserved in phosphotyrosyl phosphatases	Phosphotyrosyl phosphatase
MJ0314	MJ0398, MJ1098	PIR/JC1383 (*Pyrobaculum organotrophum*); 2.0e-11; intron and intein endonuclease signature	Endonuclease also containing a predicted DNA-binding HTH domain; MJ0314, MJ0398 and MJ1098 are newly detected, 'stand-alone' genes encoding putative endonucleases that are homologous to the 16 inteins contained in *M. jannaschii* genes (Bult *et al.*, 1996)
MJ0357	None	SECB_ECOLI; 0.036; motif conserved in SecB proteins from different bacteria	Component of protein secretion machinery
MJ0371	None	S61G_YEAST; 1.4e-05	Signal recognition particle subunit SEC61
MJ0459	None	gil496733 (*Sulfolobus solfataricus*); 7.3e-09	Translation elongation factor 1b; distantly related to eukaryotic translation elongation factors
MJ0539	Eight class I aaRSases	gil496154 (*Mycoplasma pulmonis*); 0.027; modified 'HIGH' and 'KMSKS' motifs	Cysteinyl-tRNA synthetase; this putative aaRS is unusual in that it shows only limited similarity to other class I aaRS and is somewhat more similar to methionyl-tRNA synthetases and glutamyl-tRNA synthetases; however, analysis of the entire set of *M. jannaschii* aaRS suggests specificity for Cys

the *M. jannaschii* genome (Bult *et al.*, 1996). Examples of conserved motifs in two families of *M. jannaschii* proteins, for which no function has been reported previously, are shown in Fig. 3.

The first family includes 13 proteins predicted to possess nucleotidyltransferase (NTase) activity (Fig. 3A). Only some of them showed moderate similarity to bacterial aminoglycoside-NTases (e.g. kanamycin adenylyltransferase), but all the sequences contain a prominent conserved motif, which is a signature of a large superfamily of nucleotidyltransferases including, in addition to the aminoglycoside-NTases, poly(A) polymerases, terminal nucleotidyltransferases and eukaryotic DNA polymerase beta (Holm and Sander, 1995; Yue *et al.*, 1996; Martin and Keller, 1996). A unique feature of the putative *M. jannaschii* NTases is their small size (with the exception of two larger proteins, MJ1086 and MJ0694), close to that of the N-terminal domain of the kanamycin NTase whose tertiary structure has been resolved (Sakon *et al.*, 1993). Three similar small putative NTases with a high similarity to the *M. jannaschii* NTases were identified in *Synechocystis* sp., and one putative NTase with a moderate similarity to the *M. jannaschii* proteins was found in *H. influenzae* (Fig. 3A); no functional predictions were previously available for any of these proteins.

The conserved motif is unique for NTases, thus allowing a confident prediction of this enzymatic activity for each of the *M. jannaschii* proteins belonging to the family. Beyond that, however, the sequence similarity between the archaeal proteins and nucleotidyltransferases with known specificity is insufficient to predict their actual function or the substrate(s). Several putative NTases in *M. jannaschii* are closely related to each other, suggesting a series of relatively recent duplications, which, in five cases, appeared also to have included an adjacent gene; the two genes may together comprise a novel type of mobile element (see legend to Fig. 3A).

The second novel protein superfamily, which includes

Table 3. Continued

MJ no.	Paralogues	Best functionally relevant hit from other species; *P*-value; % identity/alignment length; conserved motifs	Predicted function/activity and comments
MJ0703	MJ0411, MJ1532	HIS4_STRCO; 2.6e-17	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide (PAICAR) isomerase
MJ0817	None	gi1143229 (*Neisseria gonorrhoae*); 2.0e-15; 28%/190	Phosphatidylserine decarboxylase
MJ0839	None	PRI1_MOUSE; 1.9e-08; 27%/222	DNA primase subunit
MJ0902	None	LEP3_ECOLI; 0.044; 25%/90	Leader peptidase
MJ0973	MJ0066	CYSH_SALTY; 9.7e-12; 30%/166; pyrophosphatase-specific P-loop (PP-motif) and additional motif conserved in PAPS reductases and ATP sulphurylases	3'-phosphoadenosine-5'-phosphosulphate (PAPS) reductase
MJ1030	MJ1030, MJ1179, MJ0541, MJ1062, MJ1255, MJ0951	YGR277c (*S. cerevisiae*); 2.3e-14; 34%/142; nucleotide-binding motifs distantly related to the 'HIGH' and 'KMSKS' motifs in class I aaARSases	Nucleotidyl(cytidylyl?) transferase
MJ1034	None	SC65_YEAST; 2.1e-07; 35%/85	Signal recognition particle subunit SEC65
MJ1206	MJ1624	PIR/JC2485 (*Lactococcus lactis*); 0.00086; 26%/218; bacterial DNA primase motifs	DNA primase (bacterial DnaG homologue); MJ1624 is a small protein with more distant similarity to DnaG that, however, retains the motifs typical of bacterial primases
MJ1207	MJ1530	PAIA_BACSU; 1.4e-08; 35%/87; acetyltransferase motifs A and B	Acetyltransferase
MJ1222	MJ0544, M1057	ALG5_YEAST; 9.9e-14; 29%/255	Dolichyl phosphate mannose synthase
MJ1311	MJ0008, MJ0133, MJ1188, MJ1614	YXDH_BACSU; 5.0e-10; 35%/262	Endonuclease
MJ1318	MJ1417	gi1142617 (*B. subtilis*); 3.8e-07; 32%/101; serine protease catalytic motifs	Periplasmic serine protease; homologue of the protease domain of bacterial Lon protease without the ATPase domain
MJ1437	MJ1594	YJJG_ECOLI; 2.6e-13; 25%/218; dehalogenase-related hydrolase motifs	Hydrolase
MJ1440	MJ1087, MJ1104, MJ1427, MJ0969	KHSE_BACSU; 8.0e-07; 27%/274; kinase motifs	Homoserine kinase
MJ1646	MJ1109, MJ1366, MJ1655, MJ0204	PYR5_HUMAN; 5.3e-09; 31%/132; phosphoribosyltransferase motifs	Orotate phosphoribosyltransferase
MJ1660	MJ0487 and seven other class II aaRSases	SYFB_YEAST; 3.6e-15; 33%/165; aaRSase class II motifs	Phenylalanyl-tRNA synthetase; this gene is clearly a duplication of MJ0487; both MJ0487 and MJ1660 are most similar to Phe-aaRS from other species; however, MJ0487 also shows significant similarity to lysyl-aaRS and is likely to be specific for lysine

12 proteins from *M. jannaschii*, is brought together by a motif with two conserved histidines and a conserved aspartic acid, which appear to be diagnostic of metal-dependent hydrolase activity (Fig. 3B). Unlike the nucleotidyltransferase family, this diverse group of proteins is ubiquitous in bacterial, archaeal and eukaryotic genomes but, as the functional significance of the conserved motif has not been recognized so far, most of them have been described only as 'hypothetical proteins'. However, two of the *Synechocystis* sp. proteins belonging to this family show a highly significant similarity to eukaryotic glyoxalases, which are well-characterized Zn enzymes (Mannervik and Ridderstrom, 1993). It is of particular interest that three of the proteins in this superfamily (MJ1236, MJ0162 and MJ0047) are highly similar to subunits of the eukaryotic cleavage and polyadenylation specificity factor (CPSF) (Chanfreau *et al.*, 1996; Jenny *et al.*, 1996; Stumpf and Domdey, 1996). It remains to be determined whether or not these proteins are involved in mRNA processing in *M. jannaschii* but, regardless of this, the conserved motif defines the likely catalytic centre of the CPSF.

As shown previously for *H. influenzae*, many of the biochemical pathways in a poorly characterized bacterium may be reconstructed on the basis of a comparison with a well-characterized, related bacterium, in that case *E. coli* (Fleischmann *et al.*, 1995; Tatusov *et al.*, 1996). This task is more complicated for an archaeon as there is no representative that has been studied as thoroughly as bacterial models. An initial reconstruction of the basic metabolic pathways for *M. jannaschii* has been performed using the WIT system (http://www.cme.msu.edu/WIT/, and R. Overbeek, personal communication).

Here, we did not aim at complete reconstruction of the *M. jannaschii* biochemistry. Nevertheless, the detailed analysis of protein sequences and, in particular, the unexpectedly high degree of sequence conservation between

many archaeal and bacterial metabolic enzymes make it possible to delineate the enzymatic complement for a number of biochemical processes in *M. jannaschii* – in several cases, significantly extending the initial reconstructions (Table 4).

A

```
                                  *                **  *
Consensus                  .uuOsua+.-....    sDu-Uuu.u.-
MJ0126          32:   AIFGSYARNEQTET---SDIDILIDYYE
MJ1217          28:   AIFGSYAREEQKET---SDIDILIDYYE
MJ0128          28:   AIFGSYARNEQTEK---SDIDILVEFYE
MJ1379          28:   ALFGSYARGEQTEE---SDIDIMVEFDE
MJ1215          32:   AIFGSYARGEQKET---SDIDIMVEFYE
MJ0435          27:   SIFGSYARNEQKET---SDIDILVEFGE
MJ0141          24:   ILFGSYARGDYDEE---SDVDVLILVKE
MJ0604          25:   ILFGSYARGDYTEE---SDIDILIVGDV
MJ1547           8:   LLYGSYAKNEYTKR---SDIDICLVGVD
MJ1305          27:   ILFGSYARGTAVEY---SDVDLLVIAKN
MJ1112          51:   LLVGSSARNTNLKD---DYDIDIFVLFDK
MJ1086         144:   GVSGSLLLKLNNKN---SDIDFVIYGKD
MJ0694         196:   GKIAEAAKNSHGGELEDYDLDVIVKFAE

HI0073          31:   WAFGSRVKGKAKKY---SDLDLAIISEE

slr1241         31:   ALFGSFLRDDFDLD---NSDIDVLVSEPN
sll1504         24:   ALFGSILRPNFHSD---SDIDILIEFAP
ssl2749         27:   ALFGSTARDEAGPH---SDVDILVSFDG

KANU_STAAU      19:   GVYGSLGRQTDGPY---SDIEMMCVMST
25A6_HUMAN     391:   VRGGSTAKGTALKT---GSDADLVVFHNS
```

B

```
Consensus              .u-.UuuaH.h.DH.ouu..u
YSH1          59:   KVDILLISHFHLDHAASLPYV
CPSF-73k      62:   EIDLLLISHFHLDHCGALPWF
CPSF-100k     51:   QIDAVLLSHPDPLHLGALPYA

MJ1236       231:   DLDAVIVTHAHLDHCGFIPGL
MJ0162        45:   AVDAVIVSHAHLDHCGAIPFY
MJ0047        51:   DVDKVFISHAHLDHSGALPVL
MJ0534        57:   KLDYIISNHISPDHNECIEKL  flavoprotein
MJ0732        56:   DLDYIIVNHVEKDHSGCVDKL  flavoprotein
MJ0748        53:   KIDVIVQNHVEKDHSGALPEI  flavoprotein
MJ0861        56:   EVKAIVLSHGHLDHIGAVPKL
MJ1163        36:   GVEVIAVTHGHADHLGNAEEL
MJ0296        74:   DIDVVINTHLHYDHIENNPIF
MJ088B        10:   DIDLIINTHCHFDHTSADYLI
MJ0301        71:   SIDMIILSHNHFDHTGGLFGI
MJ0446        54:   GFDYIVLSHGHYDHCDGLKYV
MJ1502        61:   KINHIFITHLHGDHILGIPGL
MJ1629        64:   KSNVITISHYHYDHYTPFFDD
MJ1374        56:   RTNALFISHCHPDHYTDGEII

HI1274        43:   TIEAVLLTHEHDDHTQGVSAF
HI1663        48:   NLKVLLLTHGHLDHVGAANQL
HI0061       585:   VLEKLILSHDDNDHAGGASTI

MG139         73:   KVKALFITHGHFDHIGGVPYL

sll0217       89:   SLDYLIVNHTEPDHSGLIPDL
sll0550       84:   RIDYLIVSHTEPDHSGLVKDI  flavoprotein
sll0647       56:   DITDIYVSHLHSDHVGGLEYV  flavoprotein
slr0050       54:   QLTRIFITHLHGDHIFGLNGL
slr0551       67:   KIKGMVVTHGHEDHIGGIAYH
sll1019       46:   DLVTIYNTHHHGDHVGANREL  glyoxalase
sll0514       59:   TVDLVFCSHAHRDHGLGLWQF  glyoxalase
slr1259       49:   KLTFCLETHVHADHITGAGKL
sll1036       61:   VSADIFFTHSHWDHIQGFPFF
```

Limited gene orthology and non-orthologous displacement of numerous genes between archaea and bacteria

An important measure of closeness between genomes of any two species is the fraction of the genes in each of the genomes that show similarity to genes from the other genome and, more specifically, how many of these genes are orthologues. Orthologues are genes related by vertical descent from a common ancestor and responsible for the same function in different species, in contrast to paralogues, which are homologues related by duplication and having similar but not identical functions (Fitch, 1970). We have observed previously that the great majority of *H. Influenzae* genes have orthologues in *E. coli*, making the smaller gene complement of the former almost a subset of the larger gene complement of the latter (Tatusov *et al.*, 1996). Furthermore, even in the case of the phylogenetically distant *M. genitalium*, about one-half of the genes had orthologues in both *H. influenzae* and *E. coli* (Mushegian and Koonin, 1996a; Koonin and Mushegian, 1996). We delineated the sets of likely orthologues of *M. jannaschii* genes in the three bacteria with completely sequenced genomes and in yeast (Table 5). Predictably, the fraction of *M. jannaschii* genes that have bacterial orthologues is considerably lower than observed even between phylogenetically distant bacteria, with the greatest number of orthologues detected in *Synechocystis* sp., a free-living, autotrophic bacterium with a relatively large genome. The fact that, in a comparison of the *M. jannaschii* protein sequences with those encoded in individual, complete bacterial and eukaryotic genomes, orthologues were found for 25% at most, testifies to the uniqueness of the *M.*

Fig. 3. Previously uncharacterized families of *M. jannaschii* proteins and their conservation in bacteria.
A. Putative nucleotidyltransferases. The alignment, generated using the MACAW program, includes all the sequences of 'MJ-type' nucleotidyltransferases encoded in the completely sequenced bacterial and archaeal genomes as well as the kanamycin nucleotidyltransferase from *Staphylococcus aureus* (KANU_STAAU) and human 2'-5' oligoadenylate synthetase (25A6_HUMAN). The position of the first residue of each aligned segment in the respective protein sequence is indicated by a number. The consensus includes amino acid residues conserved in all sequences (upper case) and those conserved in the majority of the sequences (lower case). U indicates a bulky hydrophobic residue (I, L, V, M, F, Y, W); O indicates a small residue (G, A, S); + indicates a positively charged residue (K, R); and – indicates a negatively charged residue (D, E). Asterisks indicate residues shown to interact with ATP in the kanamycin nucleotidyltransferases (Pedersen *et al.*, 1995).
B. Putative Zn-dependent hydrolases. The alignment of the conserved motif was constructed using the CAP and MOST programs and includes all the proteins encoded in the completely sequenced bacterial and archaeal genomes, which belong to the superfamily. In addition, the sequences of the CPSF subunits from yeast (YSH1) and from humans (CPSF-73k and CPSF-100k) are shown. The designations are as in (A).

Table 4. Reconstruction of selected metabolic pathways in *Methanococcus jannaschii*.

Pathways	Bacterial genes and their functional equivalents in *M. jannaschii*[a]
Glycolysis (the downstream portion)	Triosephosphate isomerase *tpiA* (MJ1528), glyceraldehyde 3-phosphate dehydrogenase *gap* (MJ1146), 3-phosphoglycerate kinase *pgk* (MJ0641), phosphoglyceromutase *yibO* (MJ1612), enolase *eno* (MJ0232), pyruvate kinase *pyk* (MJ0108)
TCA cycle derivative	Malate dehydrogenase *mdh* (MJ1425), fumarase *fumC* (MJ1294), fumarate reductase flavoprotein *frdA* (MJ0033) and iron-sulphur protein *frdB* (MJ0092) subunits, succinyl-CoA synthetase alpha *sucD* (MJ0210) and beta *sucC* (MJ1246) subunits
Lipid biosynthesis[b]	Acetoacyl-CoA thiolase *ERG10* (MJ1549), 3-hydroxy-3-methylglutaryl-CoA synthase *HMGS* (MJ1546), 3-hydroxy-3-methylglutaryl-CoA reductase *HMG1* (MJ0705), mevalonate kinase *ERG12* (MJ1087), phosphomevalonate kinase *ERG8* (MJ1427 and/or MJ0969), diphosphomevalonate decarboxylase *ERG19* (MJ0102), isopenthenyl-diphosphate delta-isomerase (?), geranyl diphosphate synthase *ERG20* (MJ0860)
NAD biosynthesis	Aspartate oxidase (quinolinate synthetase B) *nadB* (MJ0033), quinolinate synthetase A *nadA* (MJ0407), quinolinate phosphoribosyltransferase *nadC* (MJ0493), nicotinate-nucleotide adenylyltransferase *nadD* (?), deamido-NAD:ammonia ligase (NAD synthetase) *nadE* (MJ1352)
Cysteine biosynthesis[c]	Rhodanese-like sulphurtransferase (MJ0052), serine-pyruvate aminotransferase (MJ0959)
Haem biosynthesis	Glutamyl-tRNA reductase *hemA* (MJ0143), glutamate 1-semialdehyde aminotransferase *hemI* (MJ0603), 5-aminolaevulinate dehydratase *hemB* (MJ0643), porphobilinogen deaminase *hemC* (MJ0569), uroporphyrinogen III synthetase *hemD* (MJ0994)
Cobalamin biosynthesis[d]	Uroporphyrinogen III methylase *cysG/cobA* (MJ0965), precorrin-2 methylase *cbiL/cobI* (MJ0771),?/*cobG* (?), precorrin-3B methylase *cbiH/cobJ* (MJ0813), precorrin-4 methylase *cbiF/cobM* (MJ1578), precorrin–6 A reductase *cbiJ/cobK* (MJ0552), precorrin 6B methylase *cbiE/cobL* (MJ1522), precorrin 6B decarboxylase *cbiT/cobL* (MJ0391), precorrin-8x isomerase *cbiC/cobH* (MJ0930), cobyrinic acid *a,c*-diamide synthase *cbiA/cobB* (MJ1421), cobalt insertion protein *-/cobN* (MJ0908), cob(I)alamin adenosyltransferase *cobA/cobO* (MJ1157), cobyric acid synthase *cbiP/cobQ* (MJ0484), cobinamide synthase *cbiB/cobD* (MJ1314), cobinamide kinase/cobinamide phosphate guanylyltransferase *cobU/cobP* (?), cobalamin synthase *cobS/cobV* (MJ1438), nicotinate-nucleotide: dimethylbenzimidazole phosphoribosyltransferase *cobT/cobU* (MJ1598)
Biotin biosynthesis	Pimeloyl-CoA synthetase *bioW*[e] (MJ1297), 7-keto-8-aminopelargonate synthetase *bioF* (MJ1298), 7,8-diaminopelargonate aminotransferase *bioA* (MJ1300), dethiobiotin synthetase *bioD* (MJ1299), biotin synthetase *bioB* (MJ1296), biotin-[acetyl-CoA carboxylase] holoenzyme synthetase *birA* (MJ1619, no transcription regulation domain)
Riboflavin biosynthesis	*ribD* pyrimidine deaminase domain (MJ0430 or MJ1102), *ribD* pyrimidine reductase domain (MJ0671), 3,4-dihydroxy-2-butanone-4-phosphate synthetase *ribB* (MJ0055), 6,7-dimethyl-8-ribityllumazin synthetase *ribE* (MJ0303), riboflavin synthase *ribF* (MJ1184?), FAD synthetase *FAD1* (MJ0066 and/or MJ0973)

a. Unless otherwise noted, the pathways are modelled after *E. coli* (Neidhardt *et al.*, 1996), and genes are named after the *E. coli* orthologues. The genes are listed in the order in which the reactions proceed. The proposed non-orthologous gene displacements are underlined.
b. Based on the mevalonate pathway in halophiles (Kates, 1993), the gene names are from *S. cerevisiae*; no candidate for isopenthenyl diphosphate delta-isomerase has been identified.
c. Systems for *de novo* biosynthesis of all nucleotides and amino acids, except for cysteine, from C1 compounds have been predicted in *M. jannaschii*, based on the high similarity to well-characterized bacterial enzymes (Bult *et al.*, 1996; http://www.cme.msu.edu/WIT/). Cysteine biosynthesis in *M. jannaschii* appears to occur by a pathway that is different from the common bacterial one, as *M. jannaschii* lacks homologues of *O*-acetylserine (thiol) lyase (*cysK* or *cysM*). The likely pathway of cysteine biosynthesis, however, could be predicted based on the reversal of the cysteine catabolism pathway in eukaryotes (Nagahara *et al.*, 1995).
d. This pathway is modelled after *Salmonella typhimurium* based on the similarity with *Pseudomonas denitrificans*, as in Roth *et al.* (1993), and both gene symbols are included where appropriate. Orthologues of *P. denitrificans* genes *cobG* and *cobP* are missing in *M. jannaschii*. Several candidates with each of the activities required for these reactions (Fe-S oxidoreductase, kinase and GMP transferase) are predicted in the *M. jannaschii* genome, which also encodes the orthologues of *S. typhimurium* genes, *cbiD* (MJ0022), *cbiM* (MJ1569) and *cbiG* (MJ1144), genes with unknown functions implicated in cobalamin biosynthesis.
e. Homologue of *B. subtilis* gene, *bioW*, not found in *E. coli*.

jannaschii genome as a representative of a distinct domain of life. Given the presumed affinity of archaea with eukaryotes (Woese *et al.*, 1990), it is, however, unexpected that the number of *M. jannaschii* genes that have orthologues in yeast is smaller than the number of genes with orthologues in *Synechocystis* sp. (Table 5, and see below).

Previous comparisons of the organization of orthologous genes in bacterial genomes have shown that only a few essential operons are conserved over large phylogenetic distances (Mushegian and Koonin, 1996b). The conservation of the genome organization is even lower in *M. jannaschii* and is limited mostly to ribosomal protein operons, genes for two subunits of the DNA-dependent RNA polymerase and some ion transport operons. In the comparisons of the *M. jannaschii* gene organization with that of bacteria, the longest common string of genes contained only four genes in a row, and the longest universally conserved blocks consisted of only three genes (data not shown; URL:http://www.ncbi.nlm.nih.gov/Complete_Genomes/Gene_Strings).

Analysis of the representation of different functional categories by orthologues indicates, along with analogies, major distinctions between *M. jannaschii* and bacteria. In particular, such key components of the archaeal replication machinery as the DNA polymerase, ATPases involved in replication initiation and ATP-dependent DNA ligase

Table 5. Sequence similarity and orthologous relationships between *Methanococcus jannaschii* genes and genes from other species with completely sequenced genomes.

	Number of *M. jannaschii* genes/% of the total (1731)	
Species and number	Coding for proteins with significant sequence similarity	Orthologues
H. influenzae (1703)	561 (32.4)	334 (19.3)
M. genitalium (468)	209 (12.1)	123 (7.1)
Synechocystis sp. (3168)	676 (39.1)	454 (26.2)
S. cerevisiae (5885)	544 (31.4)	331 (19.1)

only have orthologues in eukaryotes. Conversely, *M. jannaschii* encodes no orthologues for the critical proteins of the bacterial replication apparatus, namely three subunits of DNA polymerase III (DnaE, DnaX and DnaN), the principal replicative helicase (DnaB), ATPase involved in replication initiation (DnaA), NAD-dependent DNA ligase and two DNA-binding proteins (Ssb and Dbh).

DNA repair systems and chaperone-like proteins are significantly underrepresented in *M. jannaschii* compared with both bacteria and yeast (Table 2 and data not shown). Major repair systems, e.g. the bacterial UvrABC excisionase or the mismatch repair system, which is common to bacteria and eukaryotes, are missing. *M. jannaschii* encodes a number of predicted nucleases, helicases and ATPases, some of which are homologues of bacterial repair enzymes with known activity but poorly characterized physiological role, e.g. SbcC and SbcD (Table 3), but others could only be placed in the 'general prediction' category. It appears almost certain that some of these predicted enzymes belong to repair systems, but their actual roles and modes of interaction await experimental studies.

The absence of the genes for several molecular chaperones, e.g. the HSP70 (DnaK), HSP90 and HSP40 (DnaJ) families, in *M. jannaschii* is particularly striking as these proteins are universally conserved in other genomes and appeared to be indispensable for any cell (Mushegian and Koonin, 1996a). These molecular chaperones are encoded by at least some archaea (Macario *et al.*, 1991; 1993; Gupta and Singh, 1994), including the *dnaK* gene in the methanogen *Methanosarcina mazei* (Macario *et al.*, 1991). We performed a reverse search of the *M. jannaschii* protein and nucleotide sequences with the sequences of known molecular chaperones, in order to detect putative distant homologues. This allowed us to identify the most likely candidate for the GroES co-chaperonin, which has not been detected in the original report (Bult *et al.*, 1996) or in our initial analysis of the *M. jannaschii* protein sequences. The detected candidate encoded by the MJ0073 gene, even though it shows only limited similarity to GroES, contains most of the amino acid residues that are typically conserved in the GroES proteins and aligns well with all the secondary structure elements determined from the crystal structure of GroES (Hunt *et al.*, 1996; Fig. 4).

However, no other candidate chaperones were revealed even by this additional analysis. Therefore, it is likely that the genes for molecular chaperones have been lost in the phylogenetic lineage leading to methanococci. As many molecular chaperones possess ATPase activity, one may speculate that, in *M. jannaschii* (and perhaps in other methanococci), at least some of the functions of the missing chaperones could have been relegated to a unique family of putative ATPases (Koonin, 1997, and see below).

The differences in the repertoire of DNA repair proteins and molecular chaperones in *M. jannaschii* and bacteria appear to be manifestations of a general phenomenon, which we called 'non-orthologous gene displacement' (Koonin *et al.*, 1996c), whereby the same essential function is performed by non-orthologous (that is, distantly related or completely unrelated) proteins in different organisms. In a comparison of *M. genitalium* and *H. influenzae*, we found that non-orthologous displacements involved about 5% of the *M. genitalium* (the species with the smaller genome) genes (Koonin *et al.*, 1996c; Mushegian and Koonin, 1996a). In contrast, from the theoretical minimal gene set derived by comparison of the *H. influenzae* and *M. genitalium* genomes and consisting of 256 gene products (Mushegian and Koonin, 1996a), only 127 (50%) showed significant sequence similarity to *M. jannaschii* proteins, and 90 (35%) were represented by apparent orthologues. A similar level of conservation was observed when the bacterial minimal gene set was compared with the yeast genome (Koonin and Mushegian, 1996). It appears that there is massive non-orthologous displacement of essential genes between bacteria, archaea and eukaryotes.

Families of paralogues in bacteria and archaea: the same main classes but significant differences among smaller families

A significant fraction of genes in bacteria, namely about one-half in *E. coli* and about one-third in *H. influenzae*, belong to families of paralogues, i.e. genes coding for homologous proteins with related but not identical functions (Brenner *et al.*, 1995; Koonin *et al.*, 1995; Labedan and Riley, 1995; Tatusov *et al.*, 1996; reviewed by Saier, 1996). We found that 53% of the *M. jannaschii* gene products belong to 194 families of paralogues, a fraction somewhat higher than that in *H. influenzae*, a bacterium with almost the same number of genes [the fraction of *H. influenzae* genes included in families of paralogues in this study is higher than that in the previous reports (Brenner *et al.*, 1995; Tatusov *et al.*, 1996), as we included

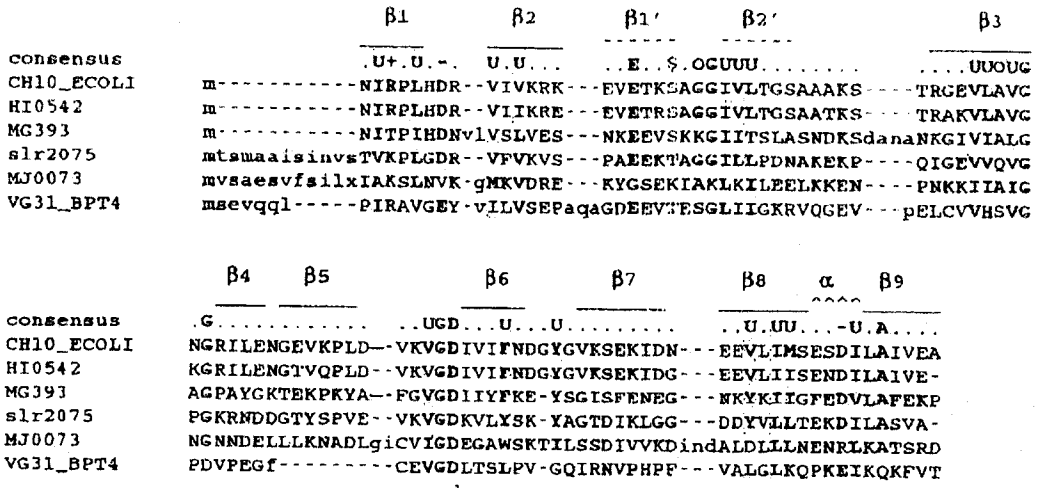

Fig. 4. The candidate GroES co-chaperonin encoded by the genome of *M. jannaschii*. In the database search with the *M. jannaschii* protein sequences, the MJ0073 sequence showed a *P*-value of 0.04 with the GroES sequence from *Porphyromonas gingivalis*, which was originally not considered significant. However, under the reciprocal procedure, when the *M. jannaschii* protein sequences were searched with the GroES sequences, MJ0073 had the lowest *P*-value of all *M. jannaschii* proteins ($< 10^{-4}$ with the *Porphyromonas gingivalis* and $< 10^{-3}$ with several other GroES sequences). The alignment was constructed using the MACAW program. The included sequences are the GroES homologues from the four completely sequenced genomes, the *E. coli* GroES protein and the gene 31 product from bacteriophage T4, which possesses a co-chaperonin activity but is only distantly related to GroES (Koonin and Van der Vlies, 1995). The consensus shows amino acid residues conserved in at least five of the six aligned sequences; the designations are as in Fig. 3. The secondary structure elements are from the crystal structure of the *E. coli* GroES (Hunt *et al.*, 1996); the dotted lines indicate the two strands that form the mobile loop directly involved in the interaction of GroES with GroEL. The 'x' in the MJ0073 sequence indicates the position of a frameshift that has been tentatively introduced in the *M. jannaschii* nucleotide sequence, resulting in an N-terminal extension of MJ073 and allowing the inclusion in the alignment of the segment corresponding to the β1 of GroES.

distant similarities detected by BLAST2 and motif analysis] and similar to that in bacteria with larger genomes, e.g. *Synechocystis* sp. and *E. coli* (Table 6; Koonin *et al.*, 1995, 1996b). Among the families of paralogues in *M. jannaschii*, 18 are unique, whereas the remaining majority is also represented in other archaea, bacteria and/or eukaryotes.

The largest classes of paralogues are the same in *M. jannaschii* and in bacteria, with the exception of the striking abundance of Fe-S oxidoreductases among the *M. jannaschii* gene products (Table 6). The expansion of this enzyme class seems to be linked to the unique biochemistry of methanococci, as many of the Fe-S oxidoreductases are involved in methanogenesis. Interestingly, ATPases and GTPases with the 'Walker-type' NTP-binding motifs, NAD(FAD)-utilizing enzymes and helix–turn–helix DNA-binding proteins comprise nearly identical fractions of the gene products in *M. jannaschii* and *H. influenzae* (Table 5). This is all the more remarkable as only a minority of the proteins in each of these superfamilies are orthologues; for example, even though *M. jannaschii* and *H. influenzae* encode nearly the same number of predicted ATPases and GTPases with the 'Walker-type' ATP-binding motifs (124 and 128 respectively), only 39 *M. jannaschii* proteins in this superfamily are represented by orthologues in *H. influenzae*. Furthermore, there are families within the largest superfamilies that are unique to *M. jannaschii*. The most striking example is a family that includes 16 putative ATPases that we designated 'MJ-type' ATPases (Table 6; Koonin, 1997) and that have originally been described as the largest unique protein family in *M. jannaschii* (Bult *et al.*, 1996).

Structural features of bacterial and archaeal proteins: M. jannaschii *encodes fewer membrane proteins and more non-globular proteins than bacteria*

We compared predicted structural features of bacterial, archaeal and yeast proteins, namely the number of proteins containing signal peptides and accordingly predicted to be secreted, those that contain predicted transmembrane helices and are likely to be integral membrane proteins and those that contain coiled-coil domains and other non-globular domains. The fraction of predicted transmembrane proteins, including those that contain multiple transmembrane helices and are likely to be transporters, is remarkably similar in all three bacteria but is somewhat lower in *M. jannaschii*. Combined with a limited number of predicted transport ATPases, this may be a reflection of a lower diversity of transport systems, which is compatible with the autotrophic lifestyle of methanococci. Compared with bacteria, the archaeal proteome is clearly enriched in proteins containing coiled-coil and other non-globular domains; remarkably, about 7% of the *M. jannaschii*

Table 6. Large protein families and superfamilies and their representation in bacterial and archaeal genomes.

Family/superfamily[a]	Number of proteins(%)				Comment
	M. jannaschii	*H. influenzae*	*M. genitalium*	*Synechocystis* sp.	
All families of paralogues	918 in 194 families (52.8%)	703 in 151 families (41.3%)	165 in 46 families (35.0%)	1775 in 322 families (56.0%)	
ATPases and GTPases with 'Walker-type' NTP-binding motifs	124 (7.1%)	128 (7.5%)	56 (12.0%)	184 (5.8%)	
Transport and repair ATPases	18 (1.0%)	42 (2.5%)	18 (3.8%)	57 (1.8%)	
'MJ'ATPases	19 (1.1%)	0	0	0	
Superfamily I helicases	2 (0.1%)	4 (0.2%)	3 (0.6%)	3 (0.1%)	
Superfamily II helicases	14 (0.8%)	12 (0.7%)	3 (0.6%)	10 (0.3%)	
'MinD-like' ATPases	13 (0.7%)	1 (0.1%)	1 (0.2%)	11 (0.3%)	
GTPases	19 (1.1%)	16 (0.9%)	14 (3.0%)	31 (1.0%)	
Fe-S oxidoreductases	88 (5.1%)	22 (1.3%)	0	51 (1.6%)	
Helix–turn–helix DNA-binding proteins and domains (primarily transcription regulators)	44 (2.5%)	49 (2.9%)	3 (0.6%)	55 (1.7%)[b]	
NAD/FAD-utilizing enzymes	42 (2.4%)	40 (2.3%)	11 (2.3%)	117 (3.7%)	
SAM-dependent methyltransferases	39 (2.2%)	23 (1.4%)	6 (1.2%)	46 (1.5%)	
NTP-utilizing enzymes with the PP-motif	18 (1%)	7 (0.4%)	4 (0.9%)	7 (0.2%)	Superfamily of NTP-utilizing enzymes with a diagnostic, modified P-loop (Bork and Koonin, 1994)
CBS domains	16 (0.9%)	5 (0.3%)	1 (0.2%)	5 (0.2%)	Conserved domain with an unknown function detected in a variety of proteins including cystathionine beta synthase and IMP dehydrogenase (Bateman, 1997)
'MJ-type' nucleotidyltransferases	12 (0.7%)	1 (0.05%)	0	3 (0.1%)	See text and Fig. 3A
Zn-dependent hydrolase superfamily I	15 (0.9%)	3 (0.2%)	1 (0.2%)	9 (0.3%)	See text and Fig. 3B
Zn(Ni)-dependent hydrolase superfamily II (including adenine deaminase and dihydroorotase)	12 (0.7%)	3 (0.2%)	0	5 (0.2%)	
Two-component system receiver domains	0	6 (0.4%)	0	83 (2.6%)	
Two-component system sensor domains (histidine kinases)	0	4 (0.2%)	0	43 (1.4%)	

a. Ordered by the abundance in *M. jannaschii*; some of the superfamilies consist of several distinct families.
b. Numerous transposases that also contain the HTH domain are not included.

proteins are predicted not to contain any globular domains (Table 7).

A large fraction of archaeal proteins shows greatest similarity to bacterial homologues and a small fraction is most similar to eukaryotic homologues

Archaea are considered to be, along with bacteria and eukaryotes, one of the three primary domains of life. Phylogenetic analysis of rRNA as well as of several proteins, primarily those involved in translation and transcription, suggested the grouping of archaea with eukaryotes as opposed to bacteria (Woese and Fox, 1977; Woese, 1987; Woese *et al.*, 1990). The root of the universal tree has been placed between the eukaryotic/archaeal and the bacterial lineages by phylogenetic analysis of universally conserved pairs of paralogues (Iwabe *et al.*, 1989; Gogarten *et al.*, 1989; 1996). It has been observed before that some archaeal proteins group with bacterial proteins and, in particular, with those from Gram-positive bacteria, in phylogenetic trees (Gupta and Golding, 1995, 1996; reviewed by Gogarten *et al.*, 1996).

The complete genome analysis, however, provides a new perspective on 'eukaryotic' and 'bacterial' genes in archaea. Given the presumed archaeal–eukaryotic association, it is striking that 44% of the *M. jannaschii* protein

Table 7. Predicted structural features of bacterial and archaeal proteins.

Predicted protein class[a]	*M. jannaschii*	*H. influenzae*	*M. genitalium*	*Synechocystis* sp.
Secreted/periplasmic proteins	98 (7.2%)	152 (8.9%)	25 (5.3%)	246 (7.8%)
Lipoproteins	16 (0.9%)	28(1.6%)	16 (3.4%)	30 (1.0%)
Transmembrane helix-containing proteins[b]	320 (18.4%)	379 (22.5%)	115 (24.6%)	800 (25.3%)
Proteins with multiple transmembrane helices[c]	153 (8.8%)	226 (13.3%)	52 (11.1%)	363 (11.6%)
Proteins with coiled-coil domains[d]	635 (36.6%)	403 (23.7%)	154(32.9%)	652 (20.6%)
Non-globular domain-containing proteins[e]	772 (44.4%)	312 (18.3%)	169 (36.1%)	851 (26.9%)
Proteins without globular domains	124 (7.2%)	32 (1.9%)	18 (3.9%)	63 (2.0%)

a. The predictions were made in the above order (from top to bottom); residues for which any feature was predicted were masked, i.e. removed from consideration for further predictions.
b. This class included proteins that contained at least one predicted transmembrane helix other than a signal peptide.
c. Proteins with at least four predicted transmembrane helices.
d. Predicted coiled-coil domains of less than 21 residues were not included.
e. Predicted non-globular domains of less than 20 residues were disregarded. Regions of 30 or fewer residues bounded by non-globular domains or by a non-globular domain and the sequence terminus were merged into the adjacent non-globular domain(s). Accordingly, proteins containing no regions longer than 30 residues between non-globular domains were considered to contain no globular domains.

sequences show a significantly higher similarity to their bacterial homologues than to eukaryotic homologues as opposed to the mere 13% that are more similar to eukaryotic homologues (Fig. 5A; the remaining 43% of the proteins either show approximately the same level of similarity to the bacterial and archaeal homologues, have only archaeal homologues or have no homologues at all). The number of *M. jannaschii* proteins that have detectable homologues *only* among bacterial proteins is also considerably greater than the number of proteins that are similar only to eukaryotic proteins (Fig. 1). Importantly, a qualitatively similar breakdown of the proteins into those with the greater similarity to bacterial homologues and those most similar to eukaryotic homologues was detected among the 275 distinct protein sequences from the archaeal genus, *Sulfolobus*, currently available in the databases (Fig. 5A). As *Methanococcus* and *Sulfolobus* belong to the two principal phylogenetic divisions of the archaea, Euryarchaeota and Crenarchaeota, respectively (Pace, 1997), it appears likely that the observed quantitative prevalence of 'bacterial' proteins is typical of all archaea.

Beyond this general balance, there is a strong contrast between different functional classes of proteins (Fig. 5A). Translation and transcription are predominantly eukaryotic, although bacterial-type transcription regulators containing the helix–turn–helix DNA-binding domain are abundant (Table 5). The replication, recombination and repair class is quantitatively dominated by 'bacterial' proteins, but these are mostly accessory proteins, such as endonucleases and DNA methylases. In contrast, the key components of the replication machinery only have orthologues in eukaryotes (see above). The protein secretion apparatus appears to be hybrid, consisting of homologues of both bacterial and eukaryotic secretion proteins. All the other functional classes, including the large group of proteins for which only a general functional prediction was possible, are dominated by 'bacterial' proteins (Fig. 5A).

Thus, the archaeal gene complement consists of a majority of genes most similar to their bacterial homologues and coding primarily for metabolic enzymes, transport systems and enzymes of cell wall biogenesis, and a minority of genes with the closest similarity to their eukaryotic counterparts, which typically encode proteins involved in genome expression.

Detailed discussion of the evolution of the three domains of life is beyond the scope of this work. Much more data and careful analysis is required for a conclusive picture to emerge, and here we only briefly discuss different scenarios that may account for the mosaic composition of the archaeal gene sets. Two fundamentally different types of explanation seem possible: (i) major variations in evolutionary rates for different groups of genes in different lineages; and (ii) genome fusion and/or horizontal gene transfer accompanied by gene loss. A rate variation scenario would posit that those groups of genes that are conserved in archaea and bacteria (e.g. the majority of metabolic enzymes) underwent a dramatic increase in the evolutionary rate in the eukaryotic lineage shortly after its separation from the other two lineages. Conversely, the group of genes that appear eukaryotic in archaea (mainly translation, transcription and replication components) should have had a phase of rapid change in the early evolution of bacteria. These hypothetical epochs of rapid change should have been brief on the evolutionary scale, as both groups of proteins are highly conserved even among deeply branching lineages within

the bacterial and the eukaryotic domain. While technically possible, such complementary evolutionary explosions appear unlikely.

Thus, a scenario including genome fusion and/or horizontal gene transfer, accompanied by gene elimination, seems to be the most realistic explanation of the observed distribution of sequence similarities among archaeal protein sequences. Strong evidence for horizontal gene transfer may come from specific relationships between archaeal genes and genes from a particular bacterial lineage, e.g. Gram-positive bacteria (Gupta and Golding, 1995; 1996). However, our analysis of the complete set of *M. jannaschii* protein sequences indicates that most of them show approximately the same level of similarity to homologues from Gram-positive and Gram-negative bacteria (Fig. 5B). Among those archaeal proteins that do show significantly higher similarity to homologues from a particular bacterial lineage, affinity with Gram-negative bacteria is more frequent (Fig. 5B and C). These preliminary observations may be most consistent with a merger between an ancestral bacterium, antedating the radiation of the main bacterial lineages, and an ancestral cell from the lineage that gave rise to the eukaryotic nucleocytoplasm, followed by differential gene loss. The subsequent evolution of archaea might have included multiple additional events of horizontal transfer of bacterial genes; hence, genes specifically related to homologues from Gram-negative bacteria, Gram-positive bacteria or cyanobacteria (Fig. 5C). This view of the probable evolutionary history of archaea invokes an obvious analogy with eukaryotes, which acquired a great number of bacterial genes as a consequence of mitochondrial and chloroplast endosymbiosis.

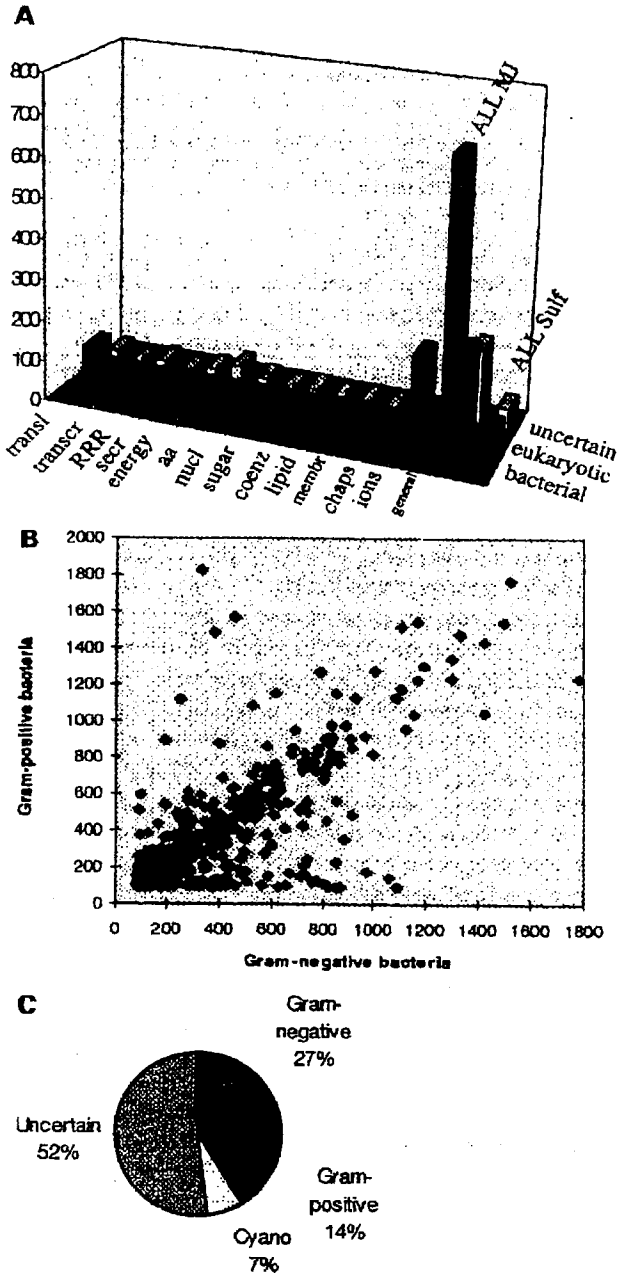

Fig. 5. Bacterial and eukaryotic homologues of archaeal proteins.
A. Proteins with highest similarity to bacterial and eukaryotic homologues in *M. jannaschii* and *Sulfolobus*
The *M. jannaschii* proteins were classified by functional category as in Table 2 (RRR, replication, recombination and repair; aa, amino acid metabolism and transport; membr, membrane biogenesis; chaps, molecular chaperones and proteins with related functions; general, proteins for which only general functional prediction was possible). For each of the archaeal proteins, the best bacterial and the best eukaryotic hit were detected using the BLATAX program, after which these hits were examined individually. A protein was considered to be significantly more similar to its bacterial homologue than to the eukaryotic homologue, or vice versa, if the difference in the percentage identity in the best alignments reported by the WUBLASTP program was at least five points and/or the difference in the reported *P*-values was at least several orders of magnitude; additionally, the conservation of domain organization was taken into account when assigning a protein to one of the classes.
B. Scatter-plot of the similarity scores between *M. jannaschii* proteins and their homologues from Gram-negative and Gram-positive bacteria. The axes show the similarity scores reported by the WUBLASTP program. The data for *M. jannaschii* proteins that had a score of at least 90 with proteins from each of the bacterial lineages are included.
C. Classification of the 'bacterial' proteins from *M. jannaschii* by sequence similarity to homologues from three major bacterial lineages. The sequences of the 745 *M. jannaschii* proteins classified as 'bacterial' were further analysed using the same criteria as in A.

Conclusions

Regardless of the phylogenetic position of the organism, sequence similarity to proteins from other species could be detected and a function could be predicted, at least at

a general level, for a large majority of the gene products. The fraction of proteins containing regions conserved over long phylogenetic distances is approximately the same in bacteria and archaea and is close to 70%. Thus, the application of sensitive methods and detailed analysis of conserved motifs makes the archaeal genomes as amenable to meaningful interpretation by computer as bacterial genomes.

The archaeal genomes encode many more proteins with a significantly higher similarity to bacterial homologues than proteins, for which the closest homologue is eukaryotic. This mosaic composition of the archaeal gene set, together with the relatively small fraction of *M. jannaschii* proteins that have orthologues among the genes in each of the other completely sequenced genomes, is compatible with the notion of archaea as a distinct domain of life. In a similar fashion to that by which eukaryotes acquired a number of bacterial genes as a consequence of mitochondrial and chloroplast endosymbiosis, the evolution of archaea probably included at least one major merger between ancestral cells from the bacterial lineage and the lineage leading to the eukaryotic nucleocytoplasm.

Experimental procedures

Nucleotide and protein sequences and databases

The nucleotide sequence of the *H. influenzae* genome was from Fleischmann *et al.* (1995), the *M. genitalium* sequence was from Fraser *et al.* (1995), the *M. jannaschii* sequence was from Bult *et al.* (1996) and the *Synechocystis* sp. sequence was from Kaneko *et al.* (1996). The gene complements of *H. influenzae* and *M. genitalium* were re-evaluated as described previously (Tatusov *et al.*, 1996; Mushegian and Koonin, 1996a), resulting in 1703 and 468 protein-coding genes respectively. There was no attempt to reassess the gene identification in *M. jannaschii* systematically, but the originally reported intergenic regions (Bult *et al.*, 1996) were compared with protein sequence databases using the BLASTX program (see below), resulting in the identification of five previously undetected genes. In addition, two groups of three genes and six pairs of genes from the originally described gene set (Bult *et al.*, 1996) were identified as originating from a single gene disrupted by frameshifts. The resulting set of *M. jannaschii* genes used for the present analysis thus consisted of 1731 genes. The gene complement of *Synechocystis* sp., which includes 3168 genes, was used as described originally (Kaneko *et al.*, 1996).

All database screening was against the protein and nucleotide versions of the non-redundant (NR) sequence database maintained at the National Center for Biotechnology Information (NIH, Bethesda, MD, USA).

The information on biochemical pathways was, in part, from the WIT database (http://www.cme.msu.edu/WIT/), the Boehringer Mannheim metabolic map (http://expasy.hcuge.ch/cgi-bin/search-biochem-index) and the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg/kegg2.html).

Database searches and protein sequence comparisons

Searches of the protein version of the NR database were performed using both the BLASTP program (Altschul *et al.*, 1990) and the WUBLASTP program based on the BLAST2 algorithm (Altschul and Gish, 1996). In the BLAST2 algorithm, the extreme value distribution statistics for a single local alignment and the sum statistics for multiple compatible alignments, which have been used originally for BLAST, are generalized to include gapped alignments, resulting in a significantly higher search sensitivity (Altschul and Gish, 1996; E. V. Koonin and A. R. Mushegian, unpublished observations). Therefore, the results of the WUBLASTP searches were generally used as the basis for assessing sequence similarity. Low-complexity regions in protein sequences, which frequently produce spurious hits in database searches, were masked before the database search using the SEG program (Wootton and Federhen, 1996). Under these conditions, the *P*-value of 0.001 or below produced by the WUBLASTP program was considered a strong indication of homology. The relevance of the search results was evaluated essentially as described previously (Koonin *et al.*, 1996b), with particular attention given to those that were associated with *P*-values greater than 0.001. The consistency of alignments between the query sequence and different database sequences was assessed using the CAP program (Tatusov *et al.*, 1994), multiple alignment analysis and visual inspection. The conservation of patterns from the PROSITE database (Bairoch, 1996) in BLAST outputs was determined using the BLA program (Tatusov and Koonin, 1994). The ECMOT collection produced in the course of the analysis of *E. coli* protein sequences (Koonin *et al.*, 1995) was used as an additional source of protein motifs. Given the non-transitivity of database searches, in cases of a small number of matches in the database or matches to sequences of uncharacterized proteins only, additional iterations of database screening using the WUBLASTP program were performed (Koonin and Tatusov, 1994). New protein motifs and multiple alignments derived from database searches were used to screen the NR database with the programs MOST (Tatusov *et al.*, 1994) and HMMER (Eddy *et al.*, 1995) respectively. An alignment of a query sequence with a database sequence was considered relevant if it had an associated *P*-value of less than 0.001 and/or contained a known or new unique motif(s).

Screening of the protein sequence database with nucleotide sequences translated in six frames in order to detect previously unidentified genes was performed using the BLASTX program (Altschul *et al.*, 1990). Conversely, nucleotide sequence databases translated in six frames were screened with protein sequences using the TBLASTN program (Altschul *et al.*, 1990).

Multiple alignments of protein sequences were constructed using the MACAW program (Schuler *et al.*, 1991).

The results of database searches were classified according to their taxonomic origin using the BLATAX program (Koonin *et al.*, 1996b).

Identification of orthologues and paralogues

A database consisting of the sequences of all gene products from *M. jannaschii*, *H. influenzae*, *M. genitalium*, *Synechocystis* sp. and *S. cerevisiae* was compared with itself using

the WUBLASTP program, and consistent groups of potential orthologues were delineated using the ROG (Rows of Ortholog Groups) program (R. L. Tatusov, E. V. Koonin and D. J. Lipman, in preparation). The putative orthologous relationships were further examined case by case in order to verify the statistical significance of the sequence similarity and the conservation of the domain organization between the candidate orthologues (Tatusov *et al.*, 1996).

Conserved strings of genes in different genomes (with possible gaps) were detected using the GENESTRING program (Tatusov *et al.*, 1996) and the produced lists of orthologues.

In order to identify clusters (families) of paralogues in each of the species, single-linkage clustering of protein sequences was initially performed using the CLUS program (Koonin *et al.*, 1996b), based on the results of BLASTP searches. A BLASTP score of 70 was chosen as the cut-off for clustering, with low-complexity regions in protein sequences masked using the SEG program. The resulting protein families were further expanded based on the results of WUBLASTP comparisons and motif analysis. The consistency of the alignment in each of the families was verified using the CAP program, multiple alignment analysis and/or visual inspection of the WUBLASTP search results.

Analysis of structural features of proteins

Signal peptides were predicted using the SIGNALP program (Nielsen *et al.*, 1997). Predicted signal peptides containing more than 35 amino acid residues were ignored. Lipoproteins were identified using the GREF program (Walker and Koonin, 1997), according to the criteria described by Sankaran *et al.* (1995). Transmembrane helices were predicted using the PHDtopology program (Rost *et al.*, 1995; Rost, 1996). Predicted helices shorter than 17 residues were disregarded. Coiled-coil regions were identified using the COILS2 program (Lupas, 1996). Non-globular domains were predicted using the SEG program with the parameters 45 (window length), 3.4 (trigger complexity) and 3.75 (extension complexity) (Wootton and Federhen, 1996). All the structural features were predicted in batch mode for complete sets of proteins from each species, and the results were automatically integrated using the UNIPRED program (Walker and Koonin, 1997).

Prediction of protein function

Protein functions were inferred by detailed inspection of the results of database searches, motif analysis, multiple alignments and structural predictions. For each protein, an attempt was made to predict the function or activity at the appropriate level of precision in order to avoid both overprediction and omission of relevant information. The database annotation attached to the protein with the highest similarity to the given query was not automatically considered applicable, even if the similarity was highly statistically significant (cf. Bork and Bairoch, 1996), and a conservative approach was adopted in general. For example, for transport proteins, the substrate was predicted only in cases when the similarity to a permease or transport ATPase with a known specificity was much higher than the similarity to proteins from other transport systems. Analogously, for such widespread enzymes with diagnostic conserved motifs as S-adenosyl methionine (SAM)-dependent methyltransferases or different classes of hydrolases, the specificity was predicted only in cases when a clear orthologue with a known specificity was available. The consistency of functional predictions for orthologues and members of families of paralogues was ensured.

In cases when a protein responsible for a particular function could not be identified in a genome on the basis of the initial sequence analysis, a reverse search procedure was applied, whereby a set of sequences of proteins with the respective function was compared with all protein sequences encoded in the given genome as well as the complete nucleotide sequence translated in six frames.

Availability of the results

The detailed results of the computer analysis of complete bacterial and archaeal genomes are available through the World Wide Web (URL:http://www.ncbi.nlm.nih.gov/Complete_Genomes).

Acknowledgements

We thank Roman Tatusov for writing many of the computer programs used in this study and indispensable help in data handling and analysis, Douglas Rand for help in sequence analysis, Burkhard Rost for providing the PHD program, Henrik Nielsen for providing the SIGNALP program, Warren Gish for help in the installation of the WUBLASTP program, Alex Bateman and Ross Overbeek for communicating their results before publication, and David Lipman, Alexey Murzin, Ross Overbeek and William Whitman for helpful discussions.

Note added in proof

After this manuscript was submitted, we became aware of a publication that reports the detection of homologues for 214 additional (compared with the original report) proteins from *M. jannaschii*, increasing the fraction of proteins with recognized similarities to 54% [Kyrpides, N.C., Olsen, G.J., Klenk, H.-P., White, O., and Woese, C.R. (1996) *Methanococcus jannaschii* genome: revisited. *Microbial Comparative Genomics* **1:** 329–338].

References

Altschul, S.F., and Gish, W. (1996) Local alignment statistics. *Methods Enzymol* **266:** 460–481.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet* **6:** 119–129.

Bairoch, A., Bucher, P., and Hofmann, K. (1996) The PROSITE database, its status in 1995. *Nucleic Acids Res* **24:** 189–196.

Bateman, A. (1997) The structure of a domain common to archaeabacteria and the homocysteinuria disease protein. *Trends Biochem Sci* **22:** 12–13.

Bork, P., and Bairoch, A. (1996) Go hunting in sequence

databases but watch out for the traps. *Trends Genet* **12:** 425–427.

Bork, P., and Koonin, E.V. (1994) A P-loop-like motif in a widespread ATP pyrophosphatase domain – implications for the evolution of sequence motifs and enzyme activity. *Proteins: Struct Funct Genet* **20:** 347–355.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992) Comprehensive computer analysis of the 182 predicted open reading frames of yeast chromosome III. *Prot Sci* **1:** 1677–1690.

Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W., and Gillevet, P.M. (1995) Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. *Mol Microbiol* **16:** 955–967.

Brenner, S.E., Hubbard, T., Murzin, A.G., and Chothia, C. (1995) Gene duplications in *H. influenzae*. *Nature* **378:** 140.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1072.

Casari, G., Andrade, M.A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C.A., Schneider, R., Tamames, J., Valencia, A., and Sander, C. (1995) Challenging times for bioinformatics. *Nature* **376:** 647–648.

Chanfreau, G., Noble, S.M., and Guthrie, C. (1996) Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor. *Science* **274:** 1511–1514.

Curnow, A.W., Ibba, M., and Söll, D. (1996) tRNA-dependent asparagine formation. *Nature* **382:** 589–590.

Eddy, S.R., Mitchison, G., and Durbin, R. (1995) Maximum discrimination models of sequence consensus. *J Comput Biol* **2:** 9–23.

Erlani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* **347:** 203–206.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99–106.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Frantz, J.D., and Gilbert, W. (1995) A novel yeast gene product, G4P1, with a specific affinity for quadruplex nucleic acids. *J Biol Chem* **270:** 20692–20697.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B. *et al.* (1996) Life with 6000 genes. *Science* **274:** 546, 563–567.

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T. *et al.* (1989) Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci USA* **86:** 6661–6665.

Gogarten, J.P., Hilario, E., and Olendzewski, L. (1996) Gene duplications and horizontal gene transfer during early evolution. In *Evolution of Microbial Life* Roberts, D. McL., Sharp, P., Alderson, G., and Collins, M. (eds). Cambridge University Press, pp. 267–292.

Golding, G.B., and Gupta, R.S. (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol Biol Evol* **12:** 1–6.

Gupta, R.S., and Golding, G.B. (1996) The origin of the eukaryotic cell. *Trends Biochem Sci* **21:** 166–171.

Gupta, R.S., and Singh, B. (1994) Phylogenetic analysis of 70 kDa heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. *Curr Biol* **4:** 1104–1114.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24:** 4420–4449.

Holm, L., and Sander, C. (1995) DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem Sci* **20:** 345–347.

Hunt, J.F., Weaver, A.J., Landry, S.J., Gierasch, L., and Deisenhofer, J. (1996) The crystal structure of the GroES co-chaperonin at 2.8 resolution. *Nature* **379:** 37–45.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* **86:** 9355–9359.

Jenny, A., Minvielle-Sebastia, L., Parker, P.J., and Keller, W. (1996) Sequence similarity between the 73-kilodalton protein of mammalian CPSF and a subunit of yeast polyadenylation factor I. *Science* **274:** 1514–1517.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3:** 109–136.

Kates, M. (1993) Membrane lipids of archaea. In *The Biochemistry of Archaea (Archaebacteria)* Kates, M. *et al.* (eds). Amsterdam: Elsevier Science Publishers, pp. 261–295.

Koonin, E.V. (1997) Evidence for a family of archaeal ATPases. *Science* **275:** 1489–1490.

Koonin, E.V., and Bork, P. (1996) Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. *Trends Biochem Sci* **21:** 128–129.

Koonin, E.V., and Mushegian, A.R. (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr Opin Genet Dev* **6:** 757–762.

Koonin, E.V., and Tatusov, R.L. (1994) Computer analysis of bacterial dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J Mol Biol* **245:** 125–132.

Koonin, E.V., and Van der Vies, S.M. (1995) Conserved sequence motifs in bacterial and bacteriophage chaperonins. *Trends Biochem Sci* **20:** 14–15.

Koonin, E.V., Bork, P., and Sander, C. (1994) Yeast chromosome III: new gene functions. *EMBO J* **13:** 493–503.

Koonin, E.V., Tatusov, R.L., and Rudd, K.E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc Natl Acad Sci USA* **92:** 11921–11925.

Koonin, E.V., Mushegian, A.R., and Rudd, K.E. (1996a) Sequencing and analysis of bacterial genomes. *Curr Biol* **6:** 404–416.

Koonin, E.V., Tatusov, R.L., and Rudd, K.E. (1996b) Genome-scale comparison of protein sequences. *Methods Enzymol* **266:** 295–322.

Koonin, E.V., Mushegian, A.R., and Bork, P. (1996c) Non-orthologous gene displacement. *Trends Genet* **12:** 334–336.

Labedan, B., and Riley, M. (1995) Widespread protein sequence similarities: origins of *Escherichia coli* genes. *J Bacteriol* **177:** 1585–1588.

Lupas, A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol* **266:** 513–525.

Macario, A.J., Dugan, C.B., Conway de Macario, E. (1991) A dnaK homologue in the archaebacterium *Methanosarcina mazei* S6. *Gene* **108:** 133–137.

Macario, A.J., Dugan, C.B., Clarens, M., and Conway de Macario, E. (1993) dnaJ in Archaea. *Nucleic Acids Res* **21:** 2773.

Mannervik, B., and Ridderstrom, M. (1993) Catalytic and molecular properties of glyoxalase I. *Biochem Soc Trans* **21:** 515–517.

Martin, G., and Keller, W. (1996) Mutational analysis of mammalian poly (A) polymerase identifies a region for primer binding and catalytic domain, homologous to the family X polymerases, and to other nucleotidyltransferases. *EMBO J* **15:** 2593–2603.

Mosyak, L., Reshetnikova, L., Goldgur, Y., Delarue, M., and Safro, M.G. (1995) Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*. *Nat Struct Biol* **2:** 537–547.

Mushegian, A.R., and Koonin, E.V. (1996a) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* **93:** 10268–10273.

Mushegian, A.R., and Koonin, E.V. (1996b) Gene order is not conserved in bacterial evolution. *Trends Genet* **12:** 289–290.

Nagahara, N., Okazaki, T., and Nishino, T. (1995) Cytosolic mercaptopyruvate sulphurtransferase is evolutionarily related to mitochondrial rhodanese. *J Biol Chem* **270:** 16230–16235.

Neidhardt, F.C., Curtiss, R., III, Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M., and Umbarger, H.E. (eds) (1996) Escherichia coli *and* Salmonella, 2nd edn. Washington, DC: American Society for Microbiology Press.

Nielsen, H., Engelbrecht, J., Brunak, S., and Von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Prot Eng* **10:** 1–6.

Ouzounis, C.A., Kyrpides, N.C., and Sander, C. (1995) Novel protein families in Archaean genomes. *Nucleic Acids Res* **23:** 565–570.

Ouzounis, C., Casari, G., Valencia, A., and Sander, C. (1996) Novelties from the complete genome of *Mycoplasma genitalium*. *Mol Microbiol* **20:** 895–900.

Pace, N.R. (1997) A molecular view of the microbial diversity and the biosphere. *Science* **276:** 734–740.

Pedersen, L.C., Benning, M.M., and Holden, H.M. (1995) Structural investigation of the antibiotic and ATP-binding sites in kanamycin nucleotidyltransferase. *Biochemistry* **34:** 13305–13311.

Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* **266:** 525–539.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) Transmembrane helices predicted at 95% accuracy. *Prot Sci* **4:** 521–533.

Roth, J.R., Lawrence, J.G., Rubenfield, M., Kieffer-Higgins, S., and Church, G.M. (1993) Characterization of the cobalamin (vitamin B_{12}) biosynthetic genes of *Salmonella typhimurium*. *J Bacteriol* **175:** 3303–3316.

Saier, M.H., Jr. (1996) Phylogenetic approaches to the identification and characterization of protein families and superfamilies. *Microb Comp Genomics* **1:** 129–150.

Sakon, J., Liao, H.H., Kanikula, A.M., Benning, M.M., Rayment, I., and Holden, H.M. (1993) Molecular structure of kanamycin nucleotidyltransferase determined to 3.0-A resolution. *Biochemistry* **32:** 11977–11984.

Sankaran, K., Gupta, S.D., and Wu, H.C. (1995) Modification of bacterial lipoproteins. *Methods Enzymol* **250:** 683–697.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., and Sander, C. (1994) GeneQuiz: a workbench for sequence analysis. In *Proceedings of the International Conference on Intelligent Systems in Molecular Biology*. Altman, R., Brutlag, D., Karp, P., Lathrop, R., and Searls, D. (eds). Menlo Park, CA: AAAI Press, pp. 348–353.

Schuler, G.D., Altschul, S.F., and Lipman, D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins: Struct Funct Genet* **9:** 180–190.

Simos, G., Segref, A., Fasiolo, F., Hellmuth, K., Shevchenko, A., Mann, M., and Hurt, E.C. (1996) The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases. *EMBO J* **15:** 5437–5448.

Strauch, M.A., Zalkin, H., and Aronson, A.I. (1988) Characterization of the glutamyl-tRNA (Gln)-to-glutaminyl-tRNA (Gln) amidotransferase reaction *of Bacillus subtilis*. *J Bacteriol* **170:** 916–920.

Stumpf, G., and Domdey, H. (1996) Dependence of yeast pre-mRNA 3′-end processing on CFT1: a sequence homologue of the mammalian AAUAAA binding factor. *Science* **274:** 1517–1520.

Tatusov, R.L., and Koonin, E.V. (1994) A simple tool to search for sequence motifs that are conserved in BLAST outputs. *Comp Appl Biosci* **10:** 457–459.

Tatusov, R.L., Altschul, S.F., and Koonin, E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA* **91:** 12091–12095.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Borodovsky, M., Hayes, W.S., Rudd, K.E., and Koonin, E.V. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison to *Escherichia coli*. *Curr Biol* **6:** 279–291.

Walker, D.R., and Koonin, E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *ISMB* **5:** 333–339.

Woese, C.R. (1987) Bacterial evolution. *Microbiol Rev* **51:** 221–271.

Woese, C.R., and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* **74:** 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proc Natl Acad Sci USA* **87:** 4576–4579.

Wootton, J.C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554–573.

Yue, D., Maizels, N., and Weiner, A.M. (1996) CCA-adding enzymes and poly (A) polymerases are all members of the same nucleotidyltransferase superfamily: characterization of the CCA-adding enzyme from the archaeal hyperthermophile *Sulfolobus shibatae*. *RNA* **2:** 895–908.